Machine Translation with a Stochastic Grammatical Channel

Dekai WU and **Hongsing** WONG
HKUST
Human Language Technology Center
Department of Computer Science
University of Science and Technology
Clear Water Bay, Hong Kong
{dekai,wong}@cs.ust.hk

Abstract

We introduce a *stochastic grammatical channel* model for machine translation, that synthesizes several desirable characteristics of both statistical and grammatical machine translation. As with the pure statistical translation model described by Wu (1996) (in which a bracketing transduction grammar models the channel), alternative hypotheses compete probabilistically, exhaustive search of the translation hypothesis space can be performed in polynomial time, and robustness heuristics arise naturally from a language-independent inversion-transduction model. However, unlike pure statistical translation models, the generated output string is guaranteed to conform to a given target grammar. The model employs only (1) a translation lexicon, (2) a context-free grammar for the target language, and (3) a bigram language model. The fact that no explicit bilingual translation rules are used makes the model easily portable to a variety of source languages. Initial experiments show that it also achieves significant speed gains over our earlier model.

1 Motivation

Speed of statistical machine translation methods has long been an issue. A step was taken by Wu (Wu, 1996) who introduced a polynomial-time algorithm for the runtime search for an optimal translation. To achieve this, Wu's method substituted a language-independent stochastic bracketing transduction grammar (SBTG) in place of the simpler word-alignment channel models reviewed in Section 2. The SBTG channel made exhaustive search possible through dynamic programming, instead of previous "stack search" heuristics. Translation accuracy was not compromised, because the SBTG is apparently flexible enough to model word-order variation (between English and Chinese) even though it eliminates large portions of the space of word alignments. The SBTG can be regarded as a model of the language-universal hypothesis that closely related arguments tend to stay together (Wu, 1995a; Wu, 1995b).

In this paper we introduce a generalization of Wu's method with the objectives of

1. increasing translation speed further,
2. improving meaning-preservation accuracy,
3. improving grammaticality of the output, and
4. seeding a natural transition toward transduction rule models,

under the constraint of

- employing no additional knowledge resources except a grammar for the target language.

To achieve these objectives, we:

- replace Wu's SBTG channel with a full stochastic inversion transduction grammar or SITG channel, discussed in Section 3, and
- (mis-)use the target language grammar as a SITG, discussed in Section 4.

In Wu's SBTG method, the burden of generating grammatical output rests mostly on the bigram language model; explicit grammatical knowledge cannot be used. As a result, output grammaticality cannot be guaranteed. The advantage is that language-dependent syntactic knowledge resources are not needed.

We relax those constraints here by assuming a good (monolingual) context-free grammar for the target language. Compared to other knowledge resources (such as transfer rules or semantic ontologies), monolingual syntactic grammars are relatively easy to acquire or construct. We use the grammar in the SITG channel, while retaining the bigram language model. The new model facilitates explicit coding of grammatical knowledge and finer control over channel probabilities. Like Wu's SBTG model, the translation hypothesis space can be exhaustively searched in polynomial time, as shown in

Section 5. The experiments discussed in Section 6 show promising results for these directions.

2 Review: Noisy Channel Model

The statistical translation model introduced by IBM (Brown et al., 1990) views translation as a noisy channel process. The underlying generative model contains a stochastic Chinese (input) sentence generator whose output is "corrupted" by the translation channel to produce English (output) sentences. Assume, as we do throughout this paper, that the input language is English and the task is to translate into Chinese. In the IBM system, the language model employs simple n-grams, while the translation model employs several sets of parameters as discussed below. Estimation of the parameters has been described elsewhere (Brown et al., 1993).

Translation is performed in the reverse direction from generation, as usual for recognition under generative models. For each English sentence $\mathbf{e}$ to be translated, the system attempts to find the Chinese sentence $\mathbf{c}*$ such that:

$$\mathbf{c}* = \underset{\mathbf{c}}{\operatorname{argmax}} \Pr(\mathbf{c}|\mathbf{e}) = \underset{\mathbf{c}}{\operatorname{argmax}} \Pr(\mathbf{e}|\mathbf{c}) \Pr(\mathbf{c}) \quad (1)$$

In the IBM model, the search for the optimal $\mathbf{c}*$ is performed using a best-first heuristic "stack search" similar to A* methods.

One of the primary obstacles to making the statistical translation approach practical is slow speed of translation, as performed in A* fashion. This price is paid for the robustness that is obtained by using very flexible language and translation models. The language model allows sentences of arbitrary order and the translation model allows arbitrary word-order permutation. No structural constraints and explicit linguistic grammars are imposed by this model.

The translation channel is characterized by two sets of parameters: translation and alignment probabilities.[1] The translation probabilities describe lexical substitution, while alignment probabilities describe word-order permutation. The key problem is that the formulation of alignment probabilities $a(i|j, V, T)$ permits the English word in position j of a length-T sentence to map to any position i of a length-V Chinese sentence. So V^T alignments are possible, yielding an exponential space with correspondingly slow search times.

3 A SITG Channel Model

The translation channel we propose is based on the recently introduced *bilingual language modeling* approach. The model employs a stochastic version of an *inversion transduction grammar* or ITG (Wu, 1995c; Wu, 1995d; Wu, 1997). This formalism was originally developed for the purpose of parallel corpus annotation, with applications for bracketing, alignment, and segmentation. Subsequently, a method was developed to use a special case of the ITG—the aforementioned BTG—for the translation task itself (Wu, 1996). The next few paragraphs briefly review the main properties of ITGs, before we describe the SITG channel.

An ITG consists of context-free productions where terminal symbols come in *couples*, for example x/y, where x is a English word and y is an Chinese translation of x, with *singletons* of the form x/ϵ or ϵ/y representing function words that are used in only one of the languages. Any parse tree thus generates both English and Chinese strings simultaneously. Thus, the tree:

(1) [I/我 [[took/拿了 [a/一 ϵ/本 book/書]$_{\text{NP}}$]$_{\text{VP}}$
[for/給 you/你]$_{\text{PP}}$]$_{\text{VP}}$]$_{\text{S}}$

produces, for example, the mutual translations:

(2) a. [我 [[拿了 [一本書]$_{\text{NP}}$]$_{\text{VP}}$ [給你]$_{\text{PP}}$]$_{\text{VP}}$]$_{\text{S}}$
b. [I [[took [a book]$_{\text{NP}}$]$_{\text{VP}}$ [for you]$_{\text{PP}}$]$_{\text{VP}}$]$_{\text{S}}$

An additional mechanism accommodates a conservative degree of word-order variation between the two languages. With each production of the grammar is associated either a *straight* orientation or an *inverted* orientation, respectively denoted as follows:

$$\begin{array}{lcl} \text{VP} & \rightarrow & [\text{VP PP}] \\ \text{VP} & \rightarrow & \langle \text{VP PP} \rangle \end{array}$$

In the case of a production with straight orientation, the right-hand-side symbols are visited left-to-right for both the English and Chinese streams. But for a production with inverted orientation, the right-hand-side symbols are visited left-to-right for English and right-to-left for Chinese. Thus, the tree:

(3) [I/我 $\langle$[took/拿了 [a/一 ϵ/本 book/書]$_{\text{NP}}$]$_{\text{VP}}$
[for/給 you/你]$_{\text{PP}}$$\rangle_{\text{VP}}$]$_{\text{S}}$

produces translations with different word order:

(4) a. [I [[took [a book]$_{\text{NP}}$]$_{\text{VP}}$ [for you]$_{\text{PP}}$]$_{\text{VP}}$]$_{\text{S}}$
b. [我 [[給你]$_{\text{PP}}$ [拿了 [一本書]$_{\text{NP}}$]$_{\text{VP}}$]$_{\text{VP}}$]$_{\text{S}}$

The surprising ability of ITGs to accommodate nearly all word-order variation between fixed-word-order languages[2] (English and Chinese in particular), has been analyzed mathematically, linguisti-

[1] Various models have been constructed by the IBM team (Brown et al., 1993). This description corresponds to one of the simplest ones, "Model 2"; search costs for the more complex models are correspondingly higher.

[2] With the exception of higher-order phenomena such as neg-raising and wh-movement.

cally, and experimentally (Wu, 1995b; Wu, 1997). Any ITG can be transformed to an equivalent binary-branching normal form.

A stochastic ITG associates a probability with each production. It follows that a SITG assigns a probability $\Pr(\mathbf{e}, \mathbf{c}, \mathbf{q})$ to all generable trees $\mathbf{q}$ and sentence-pairs. In principle it can be used as the translation channel model by normalizing with $\Pr(\mathbf{c})$ and integrating out $\Pr(\mathbf{q})$ to give $\Pr(\mathbf{e}|\mathbf{c})$ in Equation (1). In practice, a strong language model makes this unnecessary, so we can instead optimize the simpler Viterbi approximation

$$\mathbf{c}* = \operatorname*{argmax}_{\mathbf{c}} \Pr(\mathbf{e}, \mathbf{c}, \mathbf{q}) \Pr(\mathbf{c}) \qquad (2)$$

To complete the picture we add a bigram model $g_{c_{j-1}c_j} = g(c_j \mid c_{j-1})$ for the Chinese language model $\Pr(\mathbf{c})$.

This approach was used for the SBTG channel (Wu, 1996), using the language-independent *bracketing* degenerate case of the SITG:[3]

$$\begin{array}{llll}
\mathrm{A} & \overset{a^{[]}}{\rightarrow} & [\mathrm{A\ A}] & \\
\mathrm{A} & \overset{a^{\langle\rangle}}{\rightarrow} & \langle \mathrm{A\ A} \rangle & \\
\mathrm{A} & \overset{b(x,y)}{\rightarrow} & x/y & \forall x, y \text{ lexical translations} \\
\mathrm{A} & \overset{b(x,\epsilon)}{\rightarrow} & x/\epsilon & \forall x \text{ language 1 vocabulary} \\
\mathrm{A} & \overset{b(\epsilon,y)}{\rightarrow} & \epsilon/y & \forall y \text{ language 2 vocabulary}
\end{array}$$

In the proposed model, a structured language-dependent ITG is used instead.

4 A Grammatical Channel Model

Stated radically, our novel modeling thesis is that a *mirrored* version of the *target* language grammar can parse sentences of the *source* language.

Ideally, an ITG would be tailored for the desired source and target languages, enumerating the transduction patterns specific to that language pair. Constructing such an ITG, however, requires massive manual labor effort for each language pair. Instead, our approach is to take a more readily acquired monolingual context-free grammar for the target language, and use (or perhaps misuse) it in the SITG channel, by employing the three tactics described below: *production mirroring*, *part-of-speech mapping*, and *word skipping*.

In the following, keep in mind our convention that language 1 is the source (English), while language 2 is the target (Chinese).

S	→	NP VP Punc
VP	→	V NP
NP	→	N Mod N | Prn
S	→	[NP VP Punc] | ⟨Punc VP NP⟩
VP	→	[V NP] | ⟨NP V⟩
NP	→	[N Mod N] | ⟨N Mod N⟩ | [Prn]

Figure 1: An input CFG and its mirrored ITG.

4.1 Production Mirroring

The first step is to convert the monolingual Chinese CFG to a bilingual ITG. The *production mirroring* tactic simply doubles the number of productions, transforming every monolingual production into two bilingual productions,[4] one straight and one inverted, as for example in Figure 1 where the upper Chinese CFG becomes the lower ITG. The intent of the mirroring is to add enough flexibility to allow parsing of English sentences using the language 1 side of the ITG. The extra productions accommodate reversed subconstituent order in the source language's constituents, at the same time restricting the language 2 output sentence to conform the given target grammar whether straight or inverted productions are used.

The following example illustrates how production mirroring works. Consider the input sentence *He is the son of Stephen*, which can be parsed by the ITG of Figure 1 to yield the corresponding output sentence 他是史蒂芬的兒子, with the following parse tree:

(5) [[[He/他]$_{\text{Prn}}$]$_{\text{NP}}$ [[is/是]$_{\text{V}}$ [the/ϵ]$_{\text{NOISE}}$ ⟨[son/兒子]$_{\text{N}}$ [of/的]$_{\text{Mod}}$ [Stephen/史蒂芬]$_{\text{N}}$ ⟩$_{\text{NP}}$]$_{\text{VP}}$ [./。]$_{\text{Punc}}$]$_{\text{S}}$

Production mirroring produced the inverted NP constituent which was necessary to parse *son of Stephen*, i.e., ⟨son/兒子 cf/的 Stephen/史蒂芬⟩$_{\text{NP}}$.

If the target CFG is purely binary branching, then the previous theoretical and linguistic analyses (Wu, 1997) suggest that much of the requisite constituent and word order transposition may be accommodated without change to the mirrored ITG. On the other hand, if the target CFG contains productions with long right-hand-sides, then merely inverting the subconstituent order will probably be insufficient. In such cases, a more complex transformation heuristic would be needed.

Objective 3 (improving grammaticality of the output) can be directly tackled by using a tight tar-

[3] Wu (Wu, 1996) experimented with Chinese-English translation, while this paper experiments with English-Chinese translation.

[4] Except for unary productions, which yield only one bilingual production.

get grammar. To see this, consider using a mirrored Chinese CFG to parse English sentences with the language 1 side of the ITG. *Any resulting parse tree must be consistent with the original Chinese grammar.* This follows from the fact that both the straight and inverted versions of a production have language 2 (Chinese) sides identical to the original monolingual production: inverting production orientation cancels out the mirroring of the right-hand-side symbols. Thus, the output grammaticality depends directly on the tightness of the original Chinese grammar.

In principle, with this approach a single target grammar could be used for translation from any number of other (fixed word-order) source languages, so long as a translation lexicon is available for each source language.

Probabilities on the mirrored ITG cannot be reliably estimated from bilingual data without a very large parallel corpus. A straightforward approximation is to employ EM or Viterbi training on just a monolingual target language (Chinese) corpus.

4.2 Part-of-Speech Mapping

The second problem is that the part-of-speech (PoS) categories used by the target (Chinese) grammar do not correspond to the source (English) words when the source sentence is parsed. It is unlikely that any English lexicon will list Chinese parts-of-speech.

We employ a simple *part-of-speech mapping* technique that allows the PoS tag of any corresponding word in the target language (as found in the translation lexicon) to serve as a proxy for the source word's PoS. The word *view*, for example, may be tagged with the Chinese tags nc and vn, since the translation lexicon holds both view$_{\mathrm{NN}}$/意見$_{\mathrm{nc}}$ and view$_{\mathrm{VB}}$/檢視$_{\mathrm{vn}}$.

Unknown English words must be handled differently since they cannot be looked up in the translation lexicon. The English PoS tag is first found by tagging the English sentence. A set of possible corresponding Chinese PoS tags is then found by table lookup (using a small hand-constructed mapping table). For example, NN may map to nc, loc and pref, while VB may map to vi, vn, vp, vv, vs, etc. This method generates many hypotheses and should only be used as a last resort.

4.3 Word Skipping

Regardless of how constituent-order transposition is handled, some function words simply do not occur in both languages, for example Chinese aspect markers. This is the rationale for the *singletons* mentioned in Section 3.

If we create an explicit singleton hypothesis for every possible input word, the resulting search space will be too large. To recognize singletons, we instead borrow the *word-skipping* technique from speech recognition and robust parsing. As formalized in the next section, we can do this by modifying the item extension step in our chart-parser-like algorithm. When the dot of an item is on the rightmost position, we can use such constituent, a *subtree*, to extend other items. In chart parsing, the valid subtrees that can be used to extend an item are those that are located on the adjacent right of the dot position of the item and the anticipated category of the item should also be equal to that of the subtrees. If word-skipping is to be used, the valid subtrees can be located a few positions right (or, left for the item corresponding to inverted production) to the dot position of the item. In other words, words between the dot position and the start of the subtee are skipped, and considered to be singletons.

Consider Sentence 5 again. Word-skipping handled the *the* which has no Chinese counterpart. At a certain point during translation, we have the following item: VP →[is/是]$_{\mathrm{V}}$ •NP. With word-skipping, it can be extended to VP →[is/是]$_{\mathrm{V}}$ NP• by the subtree ⟨son/兒子 of/的 Stephen/史蒂芬⟩$_{\mathrm{NP}}$, even the subtree is not adjacent (but within a certain distance, see Section 5) to the dot position of the item. The *the* located on the adjacent to the dot position of the item is skipped.

Word-skipping provides us the flexibility to parse the source input by skipping possible singleton(s), if when we doing so, the source input can be parsed with the highest likelihood, and grammatical output can be produced.

5 Translation Algorithm

The translation search algorithm differs from that of Wu's SBTG model in that it handles arbitrary grammars rather than binary bracketing grammars. As such it is more similar to active chart parsing (Earley, 1970) rather than CYK parsing (Kasami, 1965; Younger, 1967). We take the standard notion of *items* (Aho and Ullman, 1972), and use the term *anticipation* to mean an item which still has symbols right of its dot. Items that don't have any symbols right of the dot are called *subtree*.

As with Wu's SBTG model, the algorithm maximizes a probabilistic objective function, Equation (2), using dynamic programming similar to that for HMM recognition (Viterbi, 1967). The presence of the bigram model in the objective function necessitates indexes in the recurrence not only on subtrees over the source English string, but also on the delimiting words of the target Chinese substrings.

The dynamic programming exploits a recursive formulation of the objective function as follows. Some notation remarks: $\mathbf{e}_{s..t}$ denotes the subsequence of English tokens $\mathbf{e}_{s+1}, \mathbf{e}_{s+2}, \ldots, \mathbf{e}_t$. We use $C(s..t)$ to denote the set of Chinese words that are translations of the English word created by taking all tokens in $\mathbf{e}_{s..t}$ together. $C(s,t)$ denotes the set of Chinese words that are translations of any of the English words anywhere within $\mathbf{e}_{s..t}$. K is the maximum number of consecutive English words that can be skipped.[5] Finally, the argmax operator is generalized to vector notation to accommodate multiple indices.

1. **Initialization**

$$\delta^0_{rstYY} = b_i(\mathbf{e}_{s..t}/Y), \quad \begin{array}{l} 0 \le s < t \le T \\ Y \in C(s..t) \\ r\ is\ Y's\ PoS \end{array}$$

2. **Recursion**

For all r, s, t, u, v such that
$$\left\{ \begin{array}{l} r \text{ is the category of a constituent spanning } s \text{ to } t \\ 0 \le s < t \le T \\ u, v \text{ are the leftmost/rightmost words of the constituent} \end{array} \right.$$

$$\begin{aligned} \delta_{rstuv} &= \max[\delta^{[]}_{rstuv}, \delta^{\langle\rangle}_{rstuv}, \delta^{0}_{rstuv}] \\ \gamma_{rstuv} &= \left\{ \begin{array}{ll} [] & \text{if } \delta^{[]}_{rstuv} > max[\delta^{\langle\rangle}_{rstuv}, \delta^{0}_{rstuv}] \\ \langle\rangle & \text{if } \delta^{\langle\rangle}_{rstuv} > max[\delta^{[]}_{rstuv}, \delta^{0}_{rstuv}] \\ 0 & \text{otherwise} \end{array} \right. \end{aligned}$$

where[6]

$$\delta^{[]}_{rstuv} = \max_{\substack{r \to [r_0 \ldots r_n] \\ s_i < t_i \le s_{i+1} \\ 0 \le s_{i+1} - t_i \le K}} a_i(r) \prod_{i=0}^{n} \delta_{r_i s_i t_i u_i v_i} \; g_{v_i u_{i+1}}$$

$$\begin{bmatrix} \tau^{[]}_{rstuv} \\ \chi^{[]}_{q_0} \\ \cdots \\ \chi^{[]}_{q_n} \end{bmatrix} = \mathop{\mathrm{argmax}}_{\substack{r \to [r_0 \ldots r_n] \\ s_i < t_i \le s_{i+1} \\ 0 \le s_{i+1} - t_i \le K}} a_i(r) \prod_{i=0}^{n} \delta_{r_i s_i t_i u_i v_i} g_{v_i u_{i+1}}$$

$$\delta^{\langle\rangle}_{rstuv} = \max_{\substack{r \to \langle r_0 \ldots r_n \rangle \\ s_i < t_i \le s_{i+1} \\ 0 \le s_{i+1} - t_i \le K}} a_i(r) \prod_{i=0}^{n} \delta_{r_i s_i t_i u_i v_i} \; g_{v_{i+1} u_i}$$

$$\begin{bmatrix} \tau^{\langle\rangle}_{rstuv} \\ \chi^{\langle\rangle}_{q_0} \\ \cdots \\ \chi^{\langle\rangle}_{q_n} \end{bmatrix} = \mathop{\mathrm{argmax}}_{\substack{r \to \langle r_0 \ldots r_n \rangle \\ s_i < t_i \le s_{i+1} \\ 0 \le s_{i+1} - t_i \le K}} a_i(r) \prod_{i=0}^{n} \delta_{r_i s_i t_i u_i v_i} g_{v_{i+1} u_i}$$

3. **Reconstruction**

Let $q_0 = (S,\ 0,\ T,\ u,\ v)$ be the optimal root. where $(u, v) = \max_{U,V \in C(0,T)} \delta_{S\ st\ U\ V}$ For any child of $q = (r, s, t, u, v)$ is given by:

$$\text{CHILD}(q, r) = \left\{ \begin{array}{lll} \tau^{[]}_q & \chi^{[]}_{r_i s_i t_i u_i v_i}, & \text{if } \gamma_q = [] \\ \tau^{\langle\rangle}_q & \chi^{\langle\rangle}_{r_i s_i t_i u_i v_i}, & \text{if } \gamma_q = \langle\rangle \\ \text{NIL} & & \text{otherwise} \end{array} \right.$$

Assuming the number of translation per word is bounded by some constant, then the maximum size of $C(s,t)$ is proportional to $t - s$. The asymptotic time complexity for our algorithm is thus bounded by $O(T^7)$. However, note that in theory the complexity upper bound rises exponentially rather than polynomially with the size of the grammar, just as for context-free parsing (Barton et al., 1987), whereas this is not a problem for Wu's SBTG algorithm. In practice, natural language grammars are usually sufficiently constrained so that speed is actually improved over the SBTG algorithm, as discussed later.

The dynamic programming is efficiently implemented by an active-chart-parser-style agenda-based algorithm, sketched as follows:

1. **Initialization** For each word in the input sentence, put a subtree with category equal to the PoS of its translation into the agenda.
2. **Recursion** Loop while agenda is not empty:
   (a) If the current item is a subtree of category X, extend existing anticipations by calling ANTICIPATIONEXTENSION. For each rule in the grammar of $Z \to XW \ldots Y$, add an initial anticipation of the form $Z \to X \bullet W \ldots Y$ and put it into the agenda. Add subtree X to the chart.
   (b) If the current item is an anticipation of the form $Z \to W \ldots \bullet X \ldots Y$ from s to t_0, find all subtrees in the chart with category X that start at position t_1 and use each subtree to extend this anticipation by calling ANTICIPATIONEXTENSION.

ANTICIPATIONEXTENSION : Assuming the subtree we found is of category X from position s_1 to t, for any anticipation of the form $Z \to W \ldots \bullet X \ldots Y$ from s_0 to $[s_1 - K, s_1]$, extend it to $Z \to W \ldots X \bullet \ldots Y$ with span from s_0 to t and add it to the agenda.

[5] In our experiments, K was set to 4

[6] $s_0 = s,\ s_n = t,\ u_0 = u,\ v_n = v,\ g_{v_n u_{n+1}} = g_{v_{n+1} u_n} = 1,\ q_i = (r_i s_i t_i u_i v_i)$

3. **Reconstruction** The output string is recursively reconstructed from the highest likelihood subtree, with category S, that span the whole input sentence.

6 Results

The grammatical channel was tested in the SILC translation system. The translation lexicon was partly constructed by training on government transcripts from the HKUST English-Chinese Parallel Bilingual Corpus, and partly entered by hand. The corpus was sentence-aligned statistically (Wu, 1994); Chinese words and collocations were extracted (Fung and Wu, 1994; Wu and Fung, 1994); then translation pairs were learned via an EM procedure (Wu and Xia, 1995). Together with hand-constructed entries, the resulting English vocabulary is approximately 9,500 words and the Chinese vocabulary is approximately 14,500 words, with a many-to-many translation mapping averaging 2.56 Chinese translations per English word. Since the lexicon's content is mixed, we approximate translation probabilities by using the unigram distribution of the target vocabulary from a small monolingual corpus. Noise still exists in the lexicon.

The Chinese grammar we used is not tight—it was written for robust parsing purposes, and as such it over-generates. Because of this we have not yet been able to conduct a fair quantitative assessment of objective 3. Our productions were constructed with reference to a standard grammar (Beijing Language and Culture Univ., 1996) and totalled 316 productions. Not all the original productions are mirrored, since some (128) are unary productions, and others are Chinese-specific lexical constructions like S $\rightarrow$ 無論 S NP 總 S, which are obviously unnecessary to handle English. About 27.7% of the non-unary Chinese productions were mirrored and the total number of productions in the final ITG is 368.

For the experiment, 222 English sentences with a maximum length of 20 words from the parallel corpus were randomly selected. Some examples of the output are shown in Figure 2. No morphological processing has been used to correct the output, and up to now we have only been testing with a bigram model trained on extremely small corpus.

With respect to objective 1 (increasing translation speed), the new model is very encouraging. Table 1 shows that over 90% of the samples can be processed within one minute by the grammatical channel model, whereas that for the SBTG channel model is about 50%. This demonstrates the stronger constraints on the search space given by the SITG.

Time (x)	SBTG Channel	Grammatical Channel
$x < 30$ secs.	15.6%	83.3%
30 secs. $< x <$ 1 min.	34.9%	7.6%
$x >$ 1 min.	49.5%	9.1%

Table 1: Translation speed.

Sentence meaning preservation	SBTG Channel	Grammatical Channel
Correct	25.9%	32.3%
Incorrect	74.1%	67.7%

Table 2: Translation accuracy.

The natural trade-off is that constraining the structure of the input decreases robustness somewhat. Approximately 13% of the test corpus could not be parsed in the grammatical channel model. As mentioned earlier, this figure is likely to vary widely depending on the characteristics of the target grammar. Of course, one can simply back off to the SBTG model when the grammatical channel rejects an input sentence.

With respect to objective 2 (improving meaning-preservation accuracy), the new model is also promising. Table 2 shows that the percentage of meaningfully translated sentences rises from 26% to 32% (ignoring the rejected cases).[7] We have judged only whether the correct meaning is conveyed by the translation, paying particular attention to word order and grammaticality, but otherwise ignoring morphological and function word choices.

7 Conclusion

Currently we are designing a tight generation-oriented Chinese grammar to replace our robust parsing-oriented grammar. We will use the new grammar to quantitatively evaluate objective 3. We are also studying complementary approaches to the English word deletion performed by word-skipping—i.e., extensions that insert Chinese words suggested by the target grammar into the output.

The framework seeds a natural transition toward pattern-based translation models (objective 4). One

[7] These accuracy rates are relatively low because these experiments are being conducted with new lexicons and grammar on a new translation direction (English-Chinese).

can post-edit the productions of a mirrored SITG more carefully and extensively than we have done in our cursory pruning, gradually transforming the original monolingual productions into a set of true transduction rule patterns. This provides a smooth evolution from a purely statistical model toward a hybrid model, as more linguistic resources become available.

We have described a new *stochastic grammatical channel* model for statistical machine translation that exhibits several nice properties in comparison with Wu's SBTG model and IBM's word alignment model. The SITG-based channel increases translation speed, improves meaning-preservation accuracy, permits tight target CFGs to be incorporated for improving output grammaticality, and suggests a natural evolution toward transduction rule models. The input CFG is adapted for use via production mirroring, part-of-speech mapping, and word-skipping. We gave a polynomial-time translation algorithm that requires only a translation lexicon, plus a CFG and bigram language model for the target language. More linguistic knowledge about the target language is employed than in pure statistical translation models, but Wu's SBTG polynomial-time bound on search cost is retained and in fact the search space can be significantly reduced by using a good grammar. Output always conforms to the given target grammar.

Input : I entirely agree with this point of view.
Output: 我全然贊成用這觀點。
Corpus: 我絕對同意這個看法。
Input : This would create a tremendous financial
burden to taxpayers in Hong Kong.
Output: 在香港對納稅人創造驚人的金融的負擔。
Corpus: 此舉對香港的納稅人會造成沉重的經濟負擔。
Input : The Government wants, and will work for, the
best education for all the children of Hong Kong.
Output: 香港那政府可能想教育工作給那最好的人
儘來自那兒童。
Corpus: 政府亦希望本港所有兒童都能接受最好的
教育，並且會致力實現這個希望。
Input : Let me repeat one simple point yet again.
Output: 我卻又循環僅指出。
Corpus: 讓我再次重覆一個簡單的要點。
Input : We are very disappointed.
Output: 我們是很失望的。
Corpus: 我們對此並不滿意。

Figure 2: Example translation outputs from the grammatical channel model.

Acknowledgments

Thanks to the SILC group members: Xuanyin Xia, Daniel Chan, Aboy Wong, Vincent Chow & James Pang.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation, and Compiling*. Prentice Hall, Englewood Cliffs, NJ.

G. Edward Barton, Robert C. Berwick, and Eric. S Ristad. 1987. *Computational Complexity and Natural Language*. MIT Press, Cambridge, MA.

Beijing Language and Culture Univ.. 1996. *Sucheng Hanyu Chuji Jiaocheng (A Short Intensive Elementary Chinese Course)*, volume 1-4. Beijing Language And Culture Univ. Press.

Peter F. Brown, John Cocke, Stephen A. DellaPietra, Vincent J. DellaPietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):29–85.

Peter F. Brown, Stephen A. DellaPietra, Vincent J. DellaPietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the Assoc. for Computing Machinery*, 13(2):94–102.

Pascale Fung and Dekai Wu. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proc. of the 2nd Annual Workshop on Very Large Corpora*, pg 69–85, Kyoto, Aug.

T. Kasami. 1965. An efficient recognition and syntax analysis algorithm for context-free languages. Technical Report AFCRL-65-758, Air Force Cambridge Research Lab., Bedford, MA.

Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

Dekai Wu and Pascale Fung. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proc. of 4th Conf. on ANLP*, pg 180–181, Stuttgart, Oct.

Dekai Wu and Xuanyin Xia. 1995. Large-scale automatic extraction of an English-Chinese lexicon. *Machine Translation*, 9(3–4):285–313.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proc. of 32nd Annual Conf. of Assoc. for Computational Linguistics*, pg 80–87, Las Cruces, Jun.

Dekai Wu. 1995a. An algorithm for simultaneously bracketing parallel texts by aligning words. In *Proc. of 33rd Annual Conf. of Assoc. for Computational Linguistics*, pg 244–251, Cambridge, MA, Jun.

Dekai Wu. 1995b. Grammarless extraction of phrasal translation examples from parallel texts. In *TMI-95, Proc. of the 6th Intń Conf. on Theoretical and Methodological Issues in Machine Translation*, volume 2, pg 354–372, Leuven, Belgium, Jul.

Dekai Wu. 1995c. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proc. of IJCAI-95, 14th Intń Joint Conf. on Artificial Intelligence*, pg 1328–1334, Montreal, Aug.

Dekai Wu. 1995d. Trainable coarse bilingual grammars for parallel text bracketing. In *Proc. of the 3rd Annual Workshop on Very Large Corpora*, pg 69–81, Cambridge, MA, Jun.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proc. of the 34th Annual Conf. of the Assoc. for Computational Linguistics*, pg 152–158, Santa Cruz, CA, Jun.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, Sept.

David H. Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208.

Machine Translation with a Stochastic Grammatical Channel
(基於合乎文法的隨機通道的機器翻譯)

Dekai WU (吳德愷) and Hongsing WONG (黃桓誠)
(香港科技大學計算機科學系)
(人類語言科技中心)
{dekai,wong}@cs.ust.hk

摘要

我們溶合了統計化及合乎文法的機器翻譯方法的優點，創出一個應用在機器翻譯的合乎文法、隨機通道模型。像 Wu(1996)所論述的純統計化機器翻譯模型(他以一個加括轉換文法模擬當中的通道)，我們的模型能做到：
(一)以概率選取最佳的假設；
(二)用多項式時間，將全部翻譯選擇逐一搜索；及
(三)善用由語言獨立、反向轉換模型中延伸出來的穩健性啟發法。
但明顯不同的是，由我們的模型翻譯出來的句子是必定合乎目標語言文法的。我們的模型只使用：
(一)翻譯字典；
(二)目標語言的上下文無關文法；及
(三)bigram 語言模型。
因我們不需要任何特定的雙語翻譯常則，所以這模型能很容易地應用在多類源語言上。初步的實驗顯示這模型比起我們較早的模型能更顯著地提高翻譯速度。